Tuneable Approximations for the Mean and Variance of the Maximum of Heterogeneous Geometrically Distributed Random Variables

Daniel R. Jeske
Department of Statistics
University of California
Riverside, CA
USA
daniel.jeske@ucr.edu

Todd Blessinger
Food and Drug Administration
Rockville, MD
USA
TBlessin@CVM.FDA.GOV

Abstract

Analysis of the maximum of *n* independent geometrically distributed random variables arises in a variety of applications in computer science and engineering. Evaluating the mean and variance of the maximum when *n* is large presents considerable computational challenges. While approximate formulae have been proposed in the case where each geometric distribution has the same probability of success, the heterogeneous case has not received any attention. We derive an epsilon-accurate approximation for both the mean and the variance in the heterogeneous case. The approximations also apply to the homogeneous case, and offer something new with their ability to tune the approximation to any desired level of accuracy. We illustrate the formulae with an application where the heterogeneous context arose quite naturally.

1. Introduction

1.1 Problem Statement

We consider the problem of evaluating the expected value of the maximum of *n* independent geometrically distributed random variables. No tractable computational form appears to exist for the case of large *n* and heterogeneous geometric random variables. Our interest in this case is motivated by the following real world engineering problem. In a wireless broadcast transmission system, a transmitter broadcasts packets from a fixed location to mobile users located at varying distance from the transmitter. The transmission protocol is to broadcast each packet for as many times as is required in order for each mobile user to successfully receive the packet. Upon successful reception of the packet, users send an acknowledgement packet back to the transmitter. Only after the transmitter has received acknowledgement packets from each mobile user will it advance to the task of transmitting the next packet in the sequence.

Let X_i denote the number of transmissions required before the i-th user successfully receives the packet being transmitted. Because each user is a different distance from the transmitter, the probability of successful reception is typically different for each user, say p_i. Assuming that successive transmission attempts are independent, X_i has a geometric distribution with parameter p_i. That is, $\Pr(X_i = x) = p_i(1-p_i)^{x-1}$, for $x \in \{1, 2, \dots\}$. The number of times the transmitter needs to send the packet is then $X_{(n)} = \underset{1 \le i \le n}{Max} X_i$. Of interest to system designers is the expected value of $X_{(n)}$, say $\mu_{(n)}$, and an understanding of how it depends upon n and $\{p_i\}_{i=1}^n$. System throughput can be measured by the fraction of non-repetitious transmissions and can be expressed as $1/\mu_{(n)}$. Characterizing system

throughput as a function of the number of users, *n*, is a common way to benchmark the performance of a system design.

In the heterogeneous case, with $q_i = 1 - p_i$, we have

$$\mu_{(n)} = \sum_i \frac{1}{1-q_i} - \sum_{i<j} \frac{1}{1-q_i q_j} + \sum_{i<j<k} \frac{1}{1-q_i q_j q_k} - \cdots + (-1)^{n+1} \frac{1}{1-\prod_{i=1}^{n} q_i} \qquad (1)$$

For small *n*, it is a trivial matter to evaluate (1) using a computer. However, when *n* is large numerical evaluation is problematic due to the alternating series of terms that can get quite large and never become negligible. For the heterogeneous case, the computational complexity is a severe problem. In addition, direct evaluation of (1) is not a scalable solution since each term of the series involves a combinatory number of terms.

1.2 Related Work

Applications related to ours have been encountered in several other domains, although often with the simplifying assumption of equal p_i values. Weiss (1962), for example, discusses a context where homogeneous redundant elements perform a task at discrete time intervals. Each element was assumed to have a constant probability of success and the expected system life is of interest. Margolin and Winokur (1967) formulate the homogeneous problem in the context of an inverse sampling scheme where X_i denotes the number of trials required for the i-th entity to succeed and $\mu_{(n)}$ represents the number of stages of sampling required for each entity to finally succeed. Flajolet and Martin (1985) discuss the homogenous problem in the context of the analysis of various database processing algorithms, including estimating the cardinality of a set. Szpankowski and Rego (1990) dealt very thoroughly with the homogeneous problem when analyzing the performance of concurrent programming methods.

It would be natural to try and utilize extreme value theory to approximate the mean of $X_{(n)}$. However, it is well known that in the homogenous case a limiting distribution does not exist [see, for example, Anderson 1970)]. Clearly the heterogeneous case would not yield to asymptotic theory since the parameter space is grows as rapidly as *n*. Szpankowski and Rego (1990) used Cauchy's residue theorem from complex variable [see, for example, Henrici (1975)] to approximate the alternating series in (1) and arrive at a large *n* approximation

$$\tilde{\mu}_{(n)} \doteq \log_Q n + \gamma / L + 0.5 \qquad (2)$$

where $Q = q^{-1}$, $L = \log Q$ and $\lambda = 0.577...$ is the Euler constant. The approximation in (2) is simple to use and *n* does not have to be that large for it to be accurate over a wide range of *p*. Letting $\sigma^2_{(n)}$ denote the variance of $X_{(n)}$, and using $\tilde{\sigma}^2_{(n)}$ to designate the homogenous case, Szpankowski and Rego also showed for large *n*

$$\tilde{\sigma}^2_{(n)} \doteq \pi^2 / (6 \ln^2 Q) + 1/12 \ . \qquad (3)$$

It's of interest to note that (3) does not depend upon *n*. Kirschenhofer and Prodinger (1993) extended the approximations of Szpankowski and Rego to cover the *d*-th order statistic, $X_{(d)}$, and Grabner and Prodinger (1997) further extended the approximation to cover the case where X_i has a negative binomial distribution.

2. Approximations

Our proposed approximations for the case of general p_i values take the form

$$\mu_{(n)} \doteq \sum_{x=0}^{k-1} [\,1 - \prod_{i=1}^{n} (1-q_i^x)\,] + \sum_{i=1}^{n} q_i^k / p_i$$

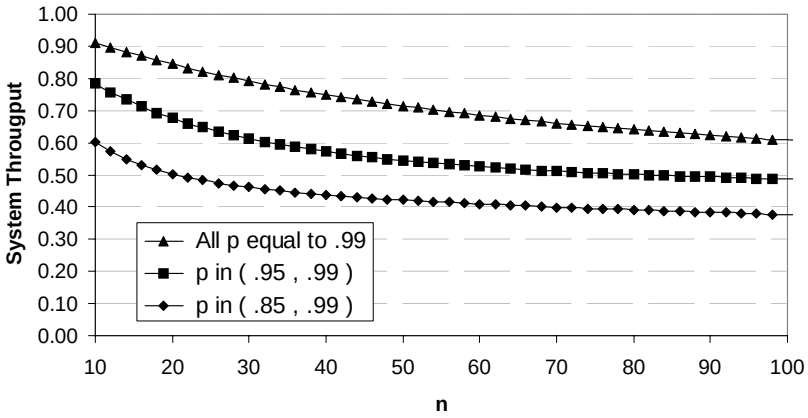
$$\sigma^2_{(n)} \doteq \sum_{x=0}^{k-1} (2x+1)[\,1 - \prod_{i=1}^{n} (1-q_i^x)\,] + \sum_{i=1}^{n} q_i^k\,[\,(2k-1)p_i + 2\,]/\,p_i^2 - \left\{ \sum_{x=0}^{k-1} [\,1 - \prod_{i=1}^{n} (1-q_i^x)\,] + \sum_{i=1}^{n} q_i^k / p_i \right\}^2 .$$

where k is selected based on prespecified ε-level of accuracy. We show that the required value of k for ε-level of accuracy of the approximation for the mean satisfies $k \geq Max\,\{-\log_{q_{(n)}}(n-2)\,,\, \log_{q_{(n)}}[(1-q_{(n)}^2)\varepsilon / C_2^n\,]^{1/2}\}$. We provide a similar bound for k that guarantees ε-level of accuracy of the approximation for the variance. In each case, the critical value of k depends on n and $q_{(n)}$. Table 1 tabulates the critical values of k for alternative combinations of $(n\,,\, q_{(n)})$.

$q_{(n)}$	n			
	20	50	100	200
0.1	3, 5	3, 6	3, 6	4, 7
0.2	4, 7	4, 8	5, 9	5, 10
0.3	5, 10	5, 11	6, 12	7, 14
0.4	6, 13	7, 15	8, 17	9, 18
0.5	8, 18	9, 20	10, 23	11, 25
0.6	11, 25	12, 29	14, 32	15, 34
0.7	15, 38	18, 43	20, 47	22, 51
0.8	25 ,66	29, 74	32, 81	35, 87
0.9	55, 158	64, 176	71, 189	77, 202

3. Application

Returning to the broadcast transmission system design problem introduced in Section 1, we utilize the approximation for $\mu_{(n)}$ to develop an insight for the performance of the system. Figure 1 shows three cases of system throughput curves, computed as $1/\mu_{(n)}$. The value of $\mu_{(n)}$ was approximated using $k = 4$. In the first case, the p_i are equally spaced on $[0.85\,,\,0.99]$, in the second case they are equally spaced on $[0.95\,,\,0.99]$ and in the third case they are all equal to 0.99. In the first case, where some of the users have relatively low probabilities of packet reception, the number of simultaneous users the system can support and still provide at least 50% throughput is 21. In the second case, where most users have relatively high packet reception probabilities, the system can support up to 40 simultaneous users and still provide at least 50% throughput. In the final case, when all users have a high probability of receiving the packet, the system can support up to 337 simultaneous users and still provide at least 50% throughput.



References

Anderson, C. W. (1970), "Extreme Value Theory for a Class of Discrete Distributions with Applications to Some Stochastic Processes, *Journal of Applied Probability*, Vol. 7, pp. 99-113.

Flajolet, P. and Martin, G. N. (1985), "Probabilistic Counting Algorithms for Data Base Applications," *Journal of Computer and System Sciences*, Vol. 31, pp. 182-209.

Grabner, P. J. and Prodinger, H. (1997), "Maximum Statistics of N Random Variables Distributed by the Negative Binomial Distribution, *Probability and Computing*, Vol. 6, pp. 179-184.

Henrici, P. (1975), Applied and Computational Complex Analysis, Vol. 2, John Wiley & Sons, New York.

Kirschenhofer, P. and Prodinger, H. (1993), "A Result in Order Statistics Related to Probabilistic Counting," *Computing*, Vol. 46, pp. 15-27.

Margolin, B. H. and Winokur, H. S. (1967), "Exact Moments of the Order Statistics of the Geometric distribution and Their Relation to Inverse Sampling and Reliability of Redundant Systems," *Journal of the American Statistical Association*, Vol. 62, pp. 915-925.

Szpankowski, W. and Rego, V. (1990), "Yet Another Application of a Binomial Recurrence: Order Statistics," *Computing*, Vol. 43, pp. 401-410.

Weiss, G. (1962), "On Certain Redundant Systems which Operate at Discrete Times," *Technometrics*, 4. p. 69-74.